Recueil planifié des données : compléments sur l'échantillonnage

Philippe Letourmy, Cirad, décembre 2017

(Établi en partie à partir d'une présentation de Statistique Canada)

Toutes les méthodes vues précédemment ont concerné des plans de sondage aléatoires. La sélection probabiliste d'un échantillon repose sur le principe de la randomisation, ou procédure de sélection aléatoire des unités dans l'échantillon.

Dans ce cas, il est possible de calculer la probabilité d'inclusion de chaque unité dans l'échantillon. Grâce à l'échantillonnage aléatoire, on peut produire des estimations fiables, de même que des estimations de l'erreur d'échantillonnage et faire des inférences au sujet de la population.

Dans la suite, nous allons voir des plans d'échantillonnage, aléatoires ou non aléatoires, très utilisés, mais pour lesquels toute inférence est basée sur le modèle des observations. Nous terminerons par une introduction aux plans aléatoires à probabilités inégales, au travers d'un exercice.

1) Un plan aléatoire : l'échantillonnage systématique

Parfois appelé échantillonnage par intervalles, l'échantillonnage systématique signifie qu'il existe un écart, ou un intervalle, entre chaque unité sélectionnée qui est incluse dans l'échantillon. Il faut suivre les étapes énumérées ci-dessous pour sélectionner un échantillon systématique.

1. Numéroter de 1 à N les unités de la base de sondage (où N est la taille de la population totale).

2. Déterminer l'intervalle d'échantillonnage (K) en divisant le nombre d'unités de la population par la taille de l'échantillon désiré. Par exemple, pour sélectionner un échantillon de 100 unités à partir d'une population de 400, il faut un intervalle

d'échantillonnage de 400 ÷ 100 = 4. K = 4, par conséquent. Il faudra sélectionner une unité sur 4 pour avoir 100 unités à l'intérieur de l'échantillon.

3. Sélectionner au hasard un nombre entre 1 et K. Ce nombre s'appelle l'origine choisie au hasard et serait le premier nombre inclus dans l'échantillon. Si l'on tire 3, la troisième unité de la base de sondage serait la première unité comprise dans l'échantillon; si l'on tire 2, le début de l'échantillon serait la deuxième unité de la base de sondage.

4. Sélectionner chaque Kième (dans ce cas, chaque 4ème) unité après ce premier nombre. L'échantillon pourrait, par exemple, se composer des unités suivantes de façon à constituer un échantillon de 100 : 3 (l'origine choisie au hasard), 7, 11, 15, 19... 395, 399 (jusqu'à N, qui est 400 dans ce cas).

On peut constater, à l'aide de l'exemple fourni ci-dessus, que dans le cas d'un échantillonnage systématique, seuls quatre échantillons possibles, qui correspondent aux quatre origines équiprobables, peuvent être sélectionnés :

1, 5, 9, 13... 393, 397

2, 6, 10, 14... 394, 398

3, 7, 11, 15... 395, 399

4, 8, 12, 16... 396, 400

Il s'agit d'un échantillonnage représentatif. Chaque membre de la population ne fait partie que de l'un des quatre échantillons et chaque échantillon a une chance égale d'être sélectionné. Cela permet de constater que chaque unité a une chance sur quatre d'être sélectionnée à l'intérieur de l'échantillon. Sa probabilité d'être sélectionnée est la même que si l'on sélectionnait un échantillon aléatoire simple de 100 unités. La principale différence tient au fait que dans le cas d'un échantillonnage aléatoire simple, toute combinaison de 100 unités aurait une chance de constituer l'échantillon, tandis que dans celui d'un échantillonnage systématique, il n'y a que quatre échantillons possibles. L'ordre de la population de la base de sondage déterminera les échantillons possibles pour l'échantillonnage systématique.

On utilise souvent cette méthode dans l'industrie, où l'on sélectionne une unité pour des essais à partir d'une chaîne de production afin de s'assurer que la machinerie et l'équipement sont d'une qualité uniforme. Un opérateur à l'intérieur d'une usine pourrait, par exemple, soumettre à un contrôle de la qualité chaque $20^{ème}$ produit sur une ligne de montage. L'opérateur pourrait choisir une origine au hasard entre les

nombres 1 et 20. Cela déterminerait le premier produit à essayer; chaque 20ème produit serait ensuite soumis à des essais.

Dans les exemples utilisés jusqu'ici, l'intervalle d'échantillonnage K était toujours un nombre entier, mais ce n'est pas toujours le cas. Par exemple, si on prélève un échantillon de 30 unités d'une population qui en compte 740, votre intervalle d'échantillonnage (ou K) sera 24,7. Dans de tels cas, on peut l'arrondir au nombre entier inférieur le plus rapproché, pour avoir une taille fixe n, ou bien accepter d'avoir une taille légèrement variable : n ou n+1.

Les avantages de l'échantillonnage systématique tiennent au fait que la sélection de l'échantillon ne peut être plus facile (vous n'obtenez qu'un seul nombre aléatoire – l'origine choisie au hasard – et le reste de l'échantillon suit automatiquement) et que l'échantillon est distribué régulièrement à l'intérieur de la population répertoriée. Le plus gros inconvénient de la méthode d'échantillonnage systématique tient au fait que les échantillons possibles risquent d'être très variables, surtout s'il existe un cycle dans le plan du mode d'ordonnancement de la population inscrite sur une liste et si ce cycle coïncide d'une quelconque façon avec l'intervalle d'échantillonnage. C'est ce que l'on peut constater dans l'exemple qui suit.

Exemple : Supposez que vous dirigez une épicerie de grande surface et que vous possédez une liste des employés de chacune de ses sections. L'épicerie est divisée entre les 10 sections suivantes : le comptoir de charcuterie, la boulangerie, les caisses, les stocks, le comptoir des viandes, les fruits et légumes, la pharmacie, le magasin de photographie, le magasin de fleurs et le nettoyage à sec. Chaque section compte 10 employés, y compris un gérant (ce qui fait 100 employés au total). Votre liste est ordonnée par section, le gérant y étant énuméré le premier et les autres employés y étant ensuite inscrits dans l'ordre décroissant d'ancienneté.

Si vous voulez sonder vos employés au sujet de leurs réflexions sur leur milieu de travail, vous pourriez choisir un petit échantillon pour répondre à vos questions. Si vous utilisiez un échantillonnage systématique et si votre intervalle d'échantillonnage était 10, vous pourriez alors ne sélectionner finalement que les gérants ou que les employés de chaque section ayant le moins d'ancienneté. Ce type d'échantillon ne vous donnerait pas un portrait complet ni approprié des réflexions des employés.

Remarque importante : le plan systématique peut être vu comme un plan en grappes, dans lequel une seule grappe est tirée. Il fournit une estimation non biaisée de la moyenne de la variable d'intérêt, mais ne peut pas donner la variance de l'estimateur, sauf à ajouter des hypothèses sur la distribution des numéros des individus (par

exemple aléatoire). Si la population est distribuée au hasard dans la base de sondage, un échantillonnage systématique devrait alors produire des résultats similaires à ceux d'un échantillonnage aléatoire simple.

2) Échantillonnage non aléatoire

La différence entre les échantillonnages aléatoire et non aléatoire tient à une hypothèse de base au sujet de la nature de la population étudiée. Dans le cas de l'échantillonnage aléatoire, chaque unité a une chance d'être sélectionnée. Dans celui de l'échantillonnage non aléatoire, on suppose que la distribution des caractéristiques à l'intérieur de la population est connue. C'est ce qui fait que le chercheur croit que n'importe quel échantillon serait représentatif et que les résultats, par conséquent, seront exacts. Pour l'échantillonnage aléatoire, la randomisation est une caractéristique du processus de sélection, plutôt qu'une hypothèse au sujet de la structure de la population.

Dans le cas de l'échantillonnage non aléatoire, puisqu'on choisit arbitrairement des unités, il n'existe aucune façon d'estimer la probabilité pour une unité quelconque d'être incluse dans l'échantillon. Également, comme la méthode en question ne fournit aucunement l'assurance que chaque unité aura une chance d'être incluse dans l'échantillon, on ne peut estimer la variabilité de l'échantillonnage ni identifier le biais possible.

On ne peut mesurer la fiabilité d'un échantillonnage non aléatoire; la seule façon de mesurer la qualité des données recueillies consiste à comparer certains des résultats de l'enquête à l'information dont on dispose au sujet de la population. Encore une fois, rien ne fournit l'assurance que les estimations ne dépasseront pas un niveau acceptable d'erreur. C’est la raison pour laquelle les statisticiens hésitent à utiliser les méthodes d'échantillonnage non aléatoire.

Malgré ces inconvénients, les méthodes d'échantillonnage non aléatoire peuvent être utiles lorsqu'on désire des commentaires descriptifs au sujet des échantillons eux-mêmes. Deuxièmement, leur utilisation prend peu de temps tout en étant plus économique et plus pratique. Il existe aussi des domaines, comme la recherche sociale appliquée, où il est impossible, ou presque, d'effectuer un échantillonnage aléatoire.

L'application de la plupart des méthodes d'échantillonnage non aléatoire exige un certain effort et une certaine organisation, mais d'autres méthodes d'échantillonnage

non aléatoire, comme l'échantillonnage de commodité, sont à l'occasion appliquées et n'exigent pas de plan d'action formel.

Voici les types les plus courants des méthodes en question.

Échantillonnage de commodité ou à l'aveuglette (à proscrire sauf exception)

On appelle parfois l'échantillonnage de commodité l'échantillonnage à l'aveuglette ou accidentel. Cet échantillonnage n'est pas normalement représentatif de la population cible, parce qu'on ne sélectionne des unités d'échantillonnage dans son cas que si on peut y avoir facilement et commodément accès.

Il arrive que monsieur ou madame Tout-le-monde utilise l'échantillonnage de commodité. Un critique gastronomique, par exemple, peut goûter plusieurs entrées ou plats principaux pour juger de la qualité et de la variété d'un menu. Les reporters des stations de télévision sont, en outre, souvent à la recherche de prétendus « interviews de gens de la rue » pour déterminer comment la population perçoit un enjeu ou une question. Dans ces deux cas, on choisit l'échantillon au hasard, sans utiliser de méthode d'enquête particulière.

L'avantage évident de la méthode, c'est qu'elle est facile à utiliser, mais la présence de biais annule énormément ce dernier. Les applications utiles sont limitées ; la technique ne peut donner des résultats exacts que lorsque la population est homogène.

Un scientifique pourrait, par exemple, utiliser cette méthode pour déterminer si un petit lac est pollué. En supposant que l'eau du lac est bien mélangée, tout échantillon donnerait de l'information identique. Un scientifique pourrait puiser de l'eau n'importe où dans le lac pour faire ses mesures. Mais cette hypothèse d’homogénéité peut être fausse !

Parmi les autres exemples d'échantillonnage de commodité, mentionnons :

•les 100 premiers clients à entrer dans un grand magasin;

•les trois premières personnes qui téléphonent à une station de radio dans le cadre d'un concours qu'elle a organisé.

Échantillonnage volontaire

Comme l'expression le laisse entendre, ce type d'échantillonnage intervient lorsque des gens offrent volontairement leurs services pour l'étude dont il est question. Il serait, par exemple, difficile et contraire à l'éthique dans le cadre d'expériences psychologiques ou d'essais de produits pharmaceutiques (de tests de médicaments) de recruter au hasard pour y participer des gens du grand public. En pareils cas, on prélève l'échantillon à partir d'un groupe de volontaires. Il arrive que la participation à une étude soit rémunérée. En échange, les volontaires acceptent la possibilité d'avoir à se prêter à des processus longs ou exigeants.

Le fait d'échantillonner des participants volontaires plutôt que la population en général peut introduire des biais marqués. Souvent, à l'occasion des sondages d'opinion, seuls les gens qui se soucient assez fortement d'une façon ou d'une autre de la question étudiée ont tendance à y répondre. La majorité silencieuse n'y répond généralement pas, ce qui entraîne un important biais sur le plan de la sélection.

Les stations de radio et de télévision ont souvent recours à des sondages par ligne ouverte pour interroger un auditoire ou un public sur ses vues. Bien souvent, on ne limite ni la fréquence ni le nombre des appels téléphoniques qu'un répondant peut effectuer en pareil cas. Une personne pourrait malheureusement, de ce fait, voter à plusieurs reprises. Il faut aussi noter que les gens qui participent à de tels sondages pourraient avoir des vues différentes de celles des gens qui ne le font pas.

Échantillonnage au jugé (à proscrire sauf exception)

On utilise la méthode d'échantillonnage au jugé lorsqu'on prélève un échantillon en se fondant sur certains jugements au sujet de l'ensemble de la population. L'hypothèse qui sous-tend son utilisation est que l'enquêteur sélectionnera des unités qui seront caractéristiques de la population. La question cruciale dans ce cas est l'objectivité : Dans quelle mesure peut-on se fier à son jugement pour en arriver à un échantillon typique? L'échantillonnage au jugé est exposé aux préjugés du chercheur et est peut-être encore davantage biaisé que l'échantillonnage de commodité ou à l'aveuglette. Étant donné que l'échantillonnage au jugé reflète toutes les idées préconçues que risque d'avoir le chercheur, il peut y avoir introduction de biais importants si ces idées sont inexactes.

Les statisticiens utilisent souvent cette méthode dans le cadre d'études préparatoires comme des tests préalables de questionnaires et des discussions en groupe. Ils

préfèrent également avoir recours à cette méthode à l'intérieur du cadre de laboratoires où le choix des sujets des expériences (comme des animaux, des êtres humains et des végétaux) reflète les croyances ou les convictions antérieures de l'enquêteur au sujet de la population.

La réduction du coût et du temps qu'exige l'acquisition de l'échantillon est un avantage de l'échantillonnage au jugé. Mais le biais peut être fort sensible, bien que pas toujours conscient.

Échantillonnage par quotas

L'échantillonnage par quotas est l'une des formes les plus courantes d'échantillonnage non aléatoire. Il s'effectue jusqu'à ce qu'un nombre précis d'unités (de quotas), pour diverses sous-populations, ait été sélectionnées. Puisqu'il n'existe aucune règle qui régirait la façon dont il faudrait s'y prendre pour remplir ces quotas, l'échantillonnage par quotas est un moyen pratique, mais sans plus, de satisfaire aux objectifs en matière de taille d'échantillon pour certaines sous-populations.

Les quotas peuvent être fondés sur des proportions de la population. Si une population, par exemple, compte 100 hommes et 100 femmes et s'il faut en prélever un échantillon de 20 personnes pour qu'elles participent à un concours de dégustation de colas, on peut vouloir diviser l'échantillon en proportions égales entre les sexes, ce qui donnerait 10 hommes et 10 femmes. On peut penser que l'échantillonnage par quotas est préférable à d'autres formes d'échantillonnage non aléatoire (comme l'échantillonnage au jugé), parce qu'il impose l'inclusion dans l'échantillon de membres de différentes sous-populations.

L'échantillonnage par quotas est un peu similaire à l'échantillonnage stratifié parce que, dans son cas également, les unités semblables sont regroupées. Toutefois, il en diffère sur le plan du mode de sélection. Dans le cas d'un échantillonnage aléatoire, on sélectionne les unités au hasard, tandis que dans celui d'un échantillonnage par quotas, on laisse habituellement à l'intervieweur le soin de déterminer qui sera échantillonné. Cela peut donner lieu à des biais de sélection. Les responsables d'études de marché utilisent donc souvent l'échantillonnage par quotas (pour des enquêtes ou des sondages téléphoniques, en particulier), plutôt que l'échantillonnage stratifié, parce qu'il est relativement peu coûteux et facile à administrer et a la propriété souhaitable de respecter les proportions de la population. L'échantillonnage par quotas camoufle toutefois des biais pouvant être significatifs.

Comme dans le cas de toutes les autres méthodes d'échantillonnage non aléatoire, il faut supposer pour l'échantillonnage par quotas que les personnes sélectionnées sont semblables à celles qu'on ne sélectionne pas, afin de formuler des inférences au sujet de la population. Des hypothèses aussi audacieuses sont rarement valables.

Exemple n° 1 : Le conseil des élèves d'une école, groupant collège et lycée, veut jauger l'opinion de ces derniers au sujet de la qualité de leurs activités parascolaires. Il décide d'interroger 100 des 1 000 élèves de l'école en utilisant comme sous-population les années d'études (c'est-à-dire ici, 5e, 4e, 3e, 2nde, 1re ou terminale). Le tableau ci-dessous fournit le nombre d'élèves.

Tableau 1. Nombre d'élèves inscrits à l'école, par année d'études

classe	Nombre d'élèves	Pourcentage des élèves (%)	Quota d'élèves à l'intérieur de l'échantillon de 100
5e	150	15	15
4e	220	22	22
3e	160	16	16
2nde	150	15	15
1re	200	20	20
terminale	120	12	12
Total	1000	100	100

Le conseil des élèves veut s'assurer que l'échantillon reflète le pourcentage d'élèves de chacune des années d'études. La formule est la suivante :

Pourcentage d'élèves en 2nde

= (nombre d'élèves de 2nde ÷ nombre total d'élèves) x 100 %

= (150 ÷ 1 000) x 100 = 15 %

Puisque 15 % des membres de la population de l'école sont en 2nde, l'échantillon devrait être constitué dans une proportion de 15 % d'élèves de 2nde. Par conséquent, la formule suivante calcule le nombre d'élèves de 2nde qui devraient être inclus dans l'échantillon :

Échantillon d'élèves de 2nde

= (15 % de 100) x 100

= 0,15 x 100 = 15 élèves

La principale différence entre l'échantillonnage stratifié et l'échantillonnage par quotas tient au fait que le premier entraînerait la sélection des élèves à l'aide d'une méthode d'échantillonnage aléatoire comme l'échantillonnage aléatoire simple ou l'échantillonnage systématique. On n'utilise pas une telle technique dans le cas de l'échantillonnage par quotas. On pourrait sélectionner les 15 élèves en choisissant les 15 premiers élèves de 2nde qui entreraient à l'école une journée donnée ou en choisissant 15 élèves dans les deux premières rangées d'une classe en particulier.

Le fait que l'échantillonnage par quotas ne respecte pas l'exigence fondamentale du hasard est le principal argument militant contre son utilisation. Certaines unités peuvent n'avoir aucune chance d'être sélectionnées, ou on risque de ne pas connaître leur chance de l'être. L'échantillon peut donc être biaisé.

Il est courant, mais il n'est pas nécessaire, que l'échantillonnage par quotas fasse appel à des procédures de sélection au hasard aux stades de départ, en grande partie de la même façon que le fait l'échantillonnage aléatoire. La première étape de l'échantillonnage à plusieurs degrés, par exemple, consisterait à sélectionner au hasard les régions géographiques. La différence se situe au niveau de la sélection des unités aux derniers stades du processus.

Dans le cas de l'échantillonnage à plusieurs degrés, les unités reposent sur des listes à jour pour ce qui est des régions sélectionnées et on sélectionne un échantillon suivant un processus aléatoire.

Dans le cas de l'échantillonnage par quotas, on indique à chaque intervieweur combien de répondants devraient être des hommes et combien d'entre eux, des femmes, de même que combien de gens devraient représenter les divers groupes d'âge. On calcule donc les quotas à partir des données dont on dispose pour la population; par conséquent, le sexe, les groupes d'âge ou d'autres variables démographiques sont représentés dans les bonnes proportions à l'intérieur des

échantillons. La qualité de l’échantillonnage repose en totalité sur la qualité de la « régression » de la variable d’intérêt sur les variables de classification déterminant les quotas. Plus les facteurs sont explicatifs, plus la méthode des quotas sera efficace.

L'échantillonnage par quotas est généralement moins coûteux que l'échantillonnage aléatoire. Il est également facile à administrer, compte tenu notamment du fait qu'on peut se passer de dresser la liste de la population entière, de sélectionner au hasard l'échantillon et d’exercer un suivi auprès des non-répondants. L'échantillonnage par quotas, qui est une méthode d'échantillonnage efficace lorsqu'on a instamment besoin d'information, peut être effectué indépendamment des bases de sondage qui existent. Il peut être la seule méthode d'échantillonnage appropriée dans bien des cas où il n'existe pas de base de sondage convenable pour la population étudiée.

3) Un plan aléatoire à probabilités inégales au travers d'un exercice

On désire estimer, à l'échelle d'un petit secteur géographique, le nombre de mètres linéaires d'archives stockées dans les mairies. Pour cela, on procède à un tirage aléatoire de 4 communes parmi les 10 de ce secteur, avec une probabilité d'inclusion proportionnelle à leur population. On appelle probabilité d'inclusion la probabilité pour la commune d'être incluse dans l'échantillon. Attention que ce n'est pas la probabilité d'un tirage individuel, mais celle que l'individu soit dans l'échantillon (la somme de ces probabilités doit être égale à n, la taille de l'échantillon).

1. Déterminer les probabilités d'inclusion dans l'échantillon, à partir des données suivantes :

Numéro de commune	Nom de la commune	Population
1	Grand'Combe	1300
2	Les Gras	1000
3	Les Combes	800
4	Les Fins	3000
5	Villers-le-Lac	4300
6	Montbenoît	500
7	Montlebon	1900
8	Arc-sous-Cicon	800
9	Gilley	1400
10	Morteau	5000

2. On appelle estimateur d'Horvitz-Thompson l'estimateur non biaisé du total d'une variable Y sur une population de N individus :

$$\hat{T} = \sum_{i \in S} \frac{y_i}{\alpha_i}$$

Où S est l'échantillon tiré aléatoirement (composé de n individus)

y_i est la valeur d'intérêt pour l'individu i

α_i est la probabilité d'inclusion de l'individu i dans l'échantillon

Démontrer que cet estimateur est non biaisé pour le total sur la population :

$$T = \sum_{i=1}^{N} Y_i$$

On utilisera le fait que la variable aléatoire $y_i = Y_i * t_i$; t_i étant la variable aléatoire égale à 1 ou 0 selon que l'individu i est dans l'échantillon ou non.

3. Estimer le métrage total des archives du canton à partir des résultats suivants :

Numéro de commune	Nom de la commune	Mètres d'archives
2	Les Gras	17
4	Les Fins	38
5	Villers-le-Lac	55
10	Morteau	70

Pourquoi échantillonner à probabilités inégales ? Un petit exemple

Soit une population de 4 entreprises A, B, C et D comptant respectivement 500, 100, 30 et 20 salariés. Admettons que l'on veuille estimer le nombre total de salariés (certes connu : 650) à partir d'un échantillon de taille 2 et comparons les 2 tirages suivants :

un sondage aléatoire simple sans remise,

un échantillonnage à probabilités inégales avec les probabilités ci-dessous :

Entreprise k	Effectif salarié X_k	Probabilité d'inclusion π_k
A	500	1
B	100	0,5
C	30	0,25
D	20	0,25

Avec le sondage aléatoire simple, il y a $C_4^2 = 6$ échantillons possibles :

Echantillon s	Probabilité de tirage p(s)	Estimation du nombre total de salariés des 4 entreprises
{A, B}	1/6	1200
{A, C}	1/6	1060
{A, D}	1/6	1040
{B, C}	1/6	260
{B, D}	1/6	240
{C, D}	1/6	100

Formule : $N*(X_1+X_2)/n$

En moyenne, on estime sans biais le vrai effectif total de 650 salariés. Mais l'estimateur est très dispersé : sa variance vaut 207 567 (CV ≅ 70%).

Avec le plan à probabilités inégales, seuls 3 échantillons sont possibles :

Echantillon s	Probabilité de tirage p(s)	Estimation du nombre total de salariés des 4 entreprises
{A, B}	0,5	700
{A, C}	0,25	620
{A, D}	0,25	580

Formule : $X_1/\pi_1 + X_2/\pi_2$

En moyenne, l'estimateur est aussi sans biais. Et sa variance est beaucoup plus faible, elle vaut 2 700 (CV ≅ 8%).